ps5

Yuanzu Chen

2023-02-20

```
library(dplyr)
```

```
##
## 载入程辑包：'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readr)
library(ggplot2)
```

Problem 1 loading data

```
gapminder <- read_delim("gapminder.csv")
```

```
## Rows: 13055 Columns: 25
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: "\t"
## chr  (6): iso3, name, iso2, region, sub-region, intermediate-region
## dbl (19): time, totalPopulation, fertilityRate, lifeExpectancy, childMortali...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
nrow(gapminder)
```

```
## [1] 13055
```

```
ncol(gapminder)
```

```
## [1] 25
```

```
head(gapminder, 3)
```

```
## # A tibble: 3 × 25
##   iso3  name  iso2  region sub-r…¹ inter…²  time total…³ ferti…⁴ lifeE…⁵ child…⁶
##   <chr> <chr> <chr> <chr>  <chr>   <chr>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 ABW   Aruba AW    Ameri… Latin … Caribb…  1960   54211    4.82    65.7      NA
## 2 ABW   Aruba AW    Ameri… Latin … Caribb…  1961   55438    4.66    66.1      NA
## 3 ABW   Aruba AW    Ameri… Latin … Caribb…  1962   56225    4.47    66.4      NA
## # … with 14 more variables: youthFemaleLiteracy <dbl>, youthMaleLiteracy <dbl>,
## #   adultLiteracy <dbl>, GDP_PC <dbl>, accessElectricity <dbl>,
## #   agriculturalLand <dbl>, agricultureTractors <dbl>, cerealProduction <dbl>,
## #   fertilizerHa <dbl>, co2 <dbl>, greenhouseGases <dbl>, co2_PC <dbl>,
## #   pm2.5_35 <dbl>, battleDeaths <dbl>, and abbreviated variable names
## #   ¹`sub-region`, ²`intermediate-region`, ³totalPopulation, ⁴fertilityRate,
## #   ⁵lifeExpectancy, ⁶childMortality
```

Problem 2 Descriptive statistics

Part 1

```
gapminder %>%
  summarize(countries = length(unique(name)), iso2codes = length(unique(iso2)), iso3codes = length
(unique(iso3)))
```

```
## # A tibble: 1 × 3
##   countries iso2codes iso3codes
##       <int>     <int>     <int>
## 1       250       249       253
```

Part 2a

```
gapminder %>%
  group_by(iso2) %>%
  summarise(count = length(unique(name)), countries = unique(name)) %>%
  arrange(desc(count))
```

```
## `summarise()` has grouped output by 'iso2'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 250 × 3
## # Groups:   iso2 [249]
##    iso2  count countries
##    <chr> <int> <chr>
##  1 <NA>      2 <NA>
##  2 <NA>      2 Namibia
##  3 AD        1 Andorra
##  4 AE        1 United Arab Emirates
##  5 AF        1 Afghanistan
##  6 AG        1 Antigua and Barbuda
##  7 AI        1 Anguilla
##  8 AL        1 Albania
##  9 AM        1 Armenia
## 10 AO        1 Angola
## # … with 240 more rows
```

Based on the data we get, Namibia is missing iso2 code

part 2b

```
gapminder %>%
  group_by(name) %>%
  summarise(counts = length(unique(iso3)), iso3code = unique(iso3)) %>%
  arrange(desc(counts))
```

```
## `summarise()` has grouped output by 'name'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 253 × 3
## # Groups:   name [250]
##    name           counts iso3code
##    <chr>           <int> <chr>
##  1 <NA>                4 CHANISL
##  2 <NA>                4 GBM
##  3 <NA>                4 KOS
##  4 <NA>                4 NLD_CURACAO
##  5 Afghanistan         1 AFG
##  6 Åland Islands       1 ALA
##  7 Albania             1 ALB
##  8 Algeria             1 DZA
##  9 American Samoa      1 ASM
## 10 Andorra             1 AND
## # … with 243 more rows
```

Based on the data we get, there are 4 iso3 code that do not have a corresponding name for it, and i did some research online that NLD_CURACAO is a place in Netherlands, and KOS is a island of Greek.

Part 3

```
min(gapminder$time, na.rm = TRUE)
```

```
## [1] 1960
```

```
max(gapminder$time, na.rm = TRUE)
```

```
## [1] 2019
```

The minimum year is 1960, and maximum year is 2019

Problem 3

Part 1

```
nrow(gapminder[is.na(gapminder$co2) == TRUE,])
```

```
## [1] 2658
```

```
nrow(gapminder[is.na(gapminder$co2_PC) == TRUE,])
```

```
## [1] 2661
```

```
gapminder %>%
  filter(is.na(co2) == TRUE) %>%
  group_by(time) %>%
  summarise(missing_co2 = length(co2)) %>%
  arrange(missing_co2) %>%
  tail(1)
```

```
## # A tibble: 1 × 2
##    time missing_co2
##   <dbl>       <int>
## 1  2019         217
```

```
gapminder %>%
  filter(is.na(co2_PC) == TRUE) %>%
  group_by(time) %>%
  summarise(missing_co2_PC = length(co2_PC)) %>%
  arrange(missing_co2_PC) %>%
  tail(1)
```

```
## # A tibble: 1 × 2
##    time missing_co2_PC
##   <dbl>          <int>
## 1  2019            217
```

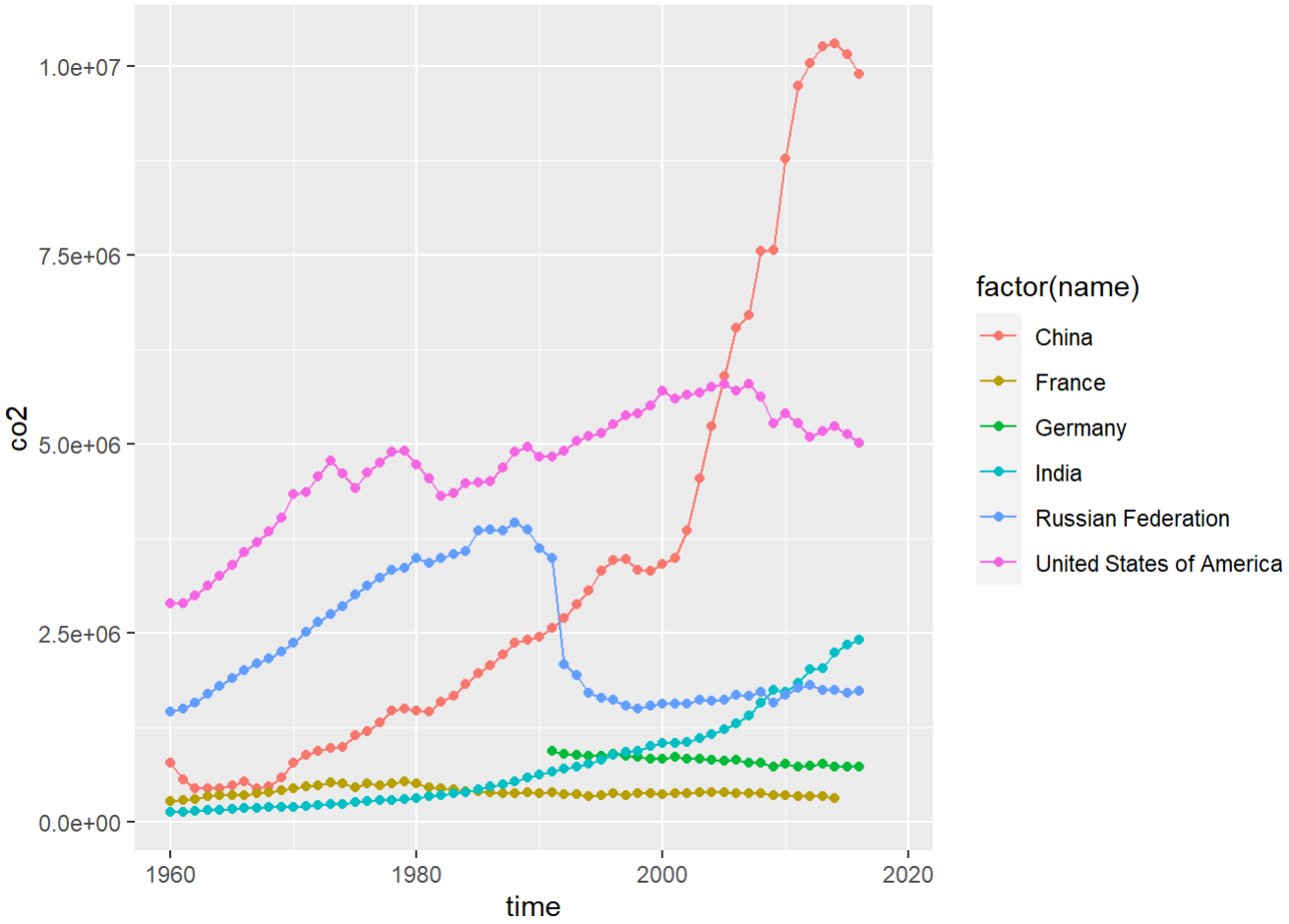
So there 2658 missing data in co2 category, and 2661 missing data in co2_PC category. 2019 has the most missing data for both co2 and co2_PC categories.

Part 2

```
plot1 <- gapminder %>%
  filter(name == "China" | name == "United States of America" | name == "India" | name == "German
y" | name == "France" | name == "Russian Federation")
ggplot(plot1, aes(time,
                  co2,
                  col = factor(name))) +
  geom_line() +
  geom_point()
```

```
## Warning: Removed 51 rows containing missing values (`geom_line()`).
```

```
## Warning: Removed 51 rows containing missing values (`geom_point()`).
```



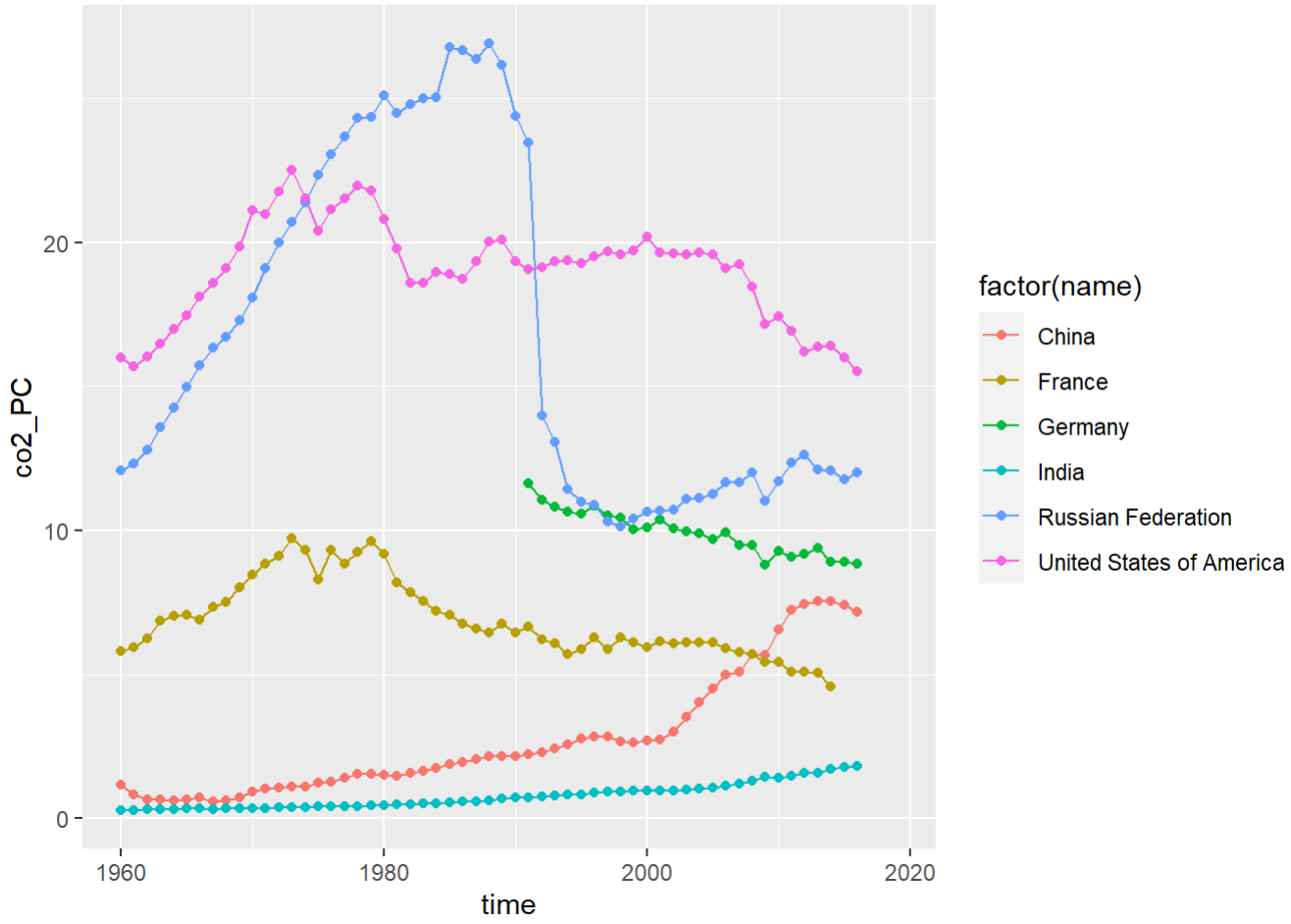
From 1970, the emission of CO2 from China is gradually increase until around 2014, it started to decrease. The emission of France and India is very constant each year. The emission of US increase from 1960 to 1973, then it drops, increase again from 1975 to 1978, then drops again, then started to slowly increase from 1980 to 2005, then it started to drop again. The emission of Russia increase from 1960 to 1987, then decrease dramatically and stay constant from 1992 to 2019.

Part 3

```
plot1 <- gapminder %>%
  filter(name == "China" | name == "United States of America" | name == "India" | name == "German
y" | name == "France" | name == "Russian Federation")
ggplot(plot1, aes(time,
                  co2_PC,
                  col = factor(name))) +
  geom_line() +
  geom_point()
```

```
## Warning: Removed 51 rows containing missing values (`geom_line()`).
```

```
## Warning: Removed 51 rows containing missing values (`geom_point()`).
```



The pattern of China, US, Germany, and Russia does not change a lot, which suggest that there overall co2 emission have some relationship with their capita. But France and India is different, France increase from 1960 to 1973, and decrease from 1979 to 2019. Whereas India's capita decrease from 1991 to 2019.

Part 4

```
co2_pc_accross_continent <- gapminder %>%
  filter(is.na(co2_PC) == FALSE) %>%
  filter(is.na(name)== FALSE)  %>%
  filter(is.na(region) == FALSE) %>%
  group_by(region, time) %>%
  summarize(average= mean(co2_PC))
```

```
## `summarise()` has grouped output by 'region'. You can override using the
## `.groups` argument.
```

```
co2_pc_accross_continent
```

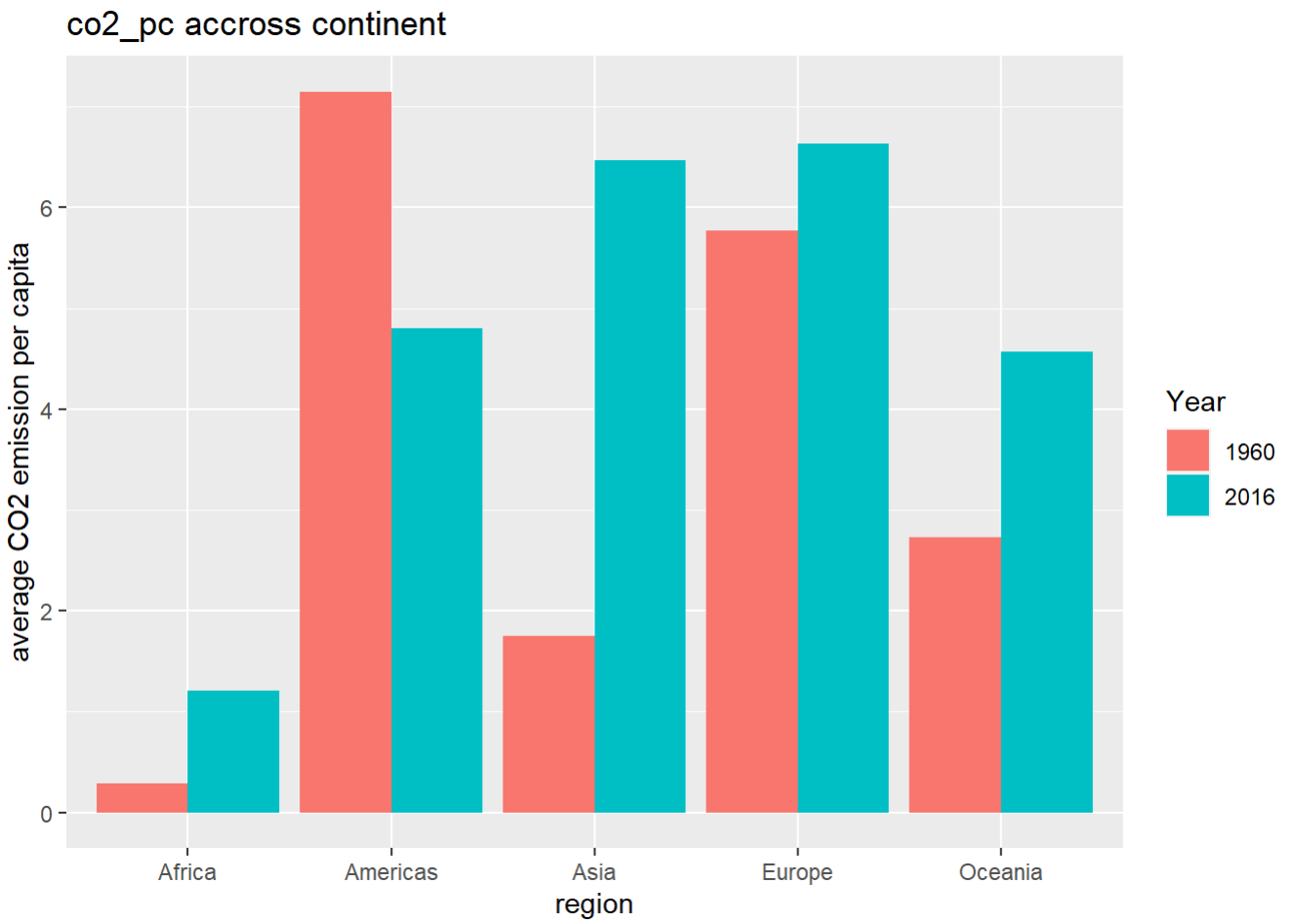
```
## # A tibble: 285 × 3
## # Groups:   region [5]
##    region  time average
##    <chr>  <dbl>   <dbl>
##  1 Africa  1960   0.291
##  2 Africa  1961   0.300
##  3 Africa  1962   0.299
##  4 Africa  1963   0.310
##  5 Africa  1964   0.349
##  6 Africa  1965   0.385
##  7 Africa  1966   0.422
##  8 Africa  1967   0.607
##  9 Africa  1968   0.781
## 10 Africa  1969   0.824
## # … with 275 more rows
```

Asia and Europe have the highest co2 emission per capita, and Africa have the least average co2 emission per capita in 2016. But in 1960, Americas have the highest emission per capita

Part 5

```
gapminder %>%
  filter(time %in% c(1960, 2016)) %>%
  filter(is.na(region) == FALSE) %>%
  filter(is.na(co2_PC) == FALSE) %>%
  filter(is.na(name) == FALSE) %>%
  group_by(region, time) %>%
  summarise(average_co2_PC = mean(co2_PC)) %>%
  ggplot(aes(region,
             average_co2_PC,
             fill = as.factor(time))) +
  geom_col(position = "dodge") +
  scale_fill_discrete("Year") +
  labs(title = "co2_pc accross continent", x = "region", y = "average CO2 emission per capita")
```

```
## `summarise()` has grouped output by 'region'. You can override using the
## `.groups` argument.
```



Part 6

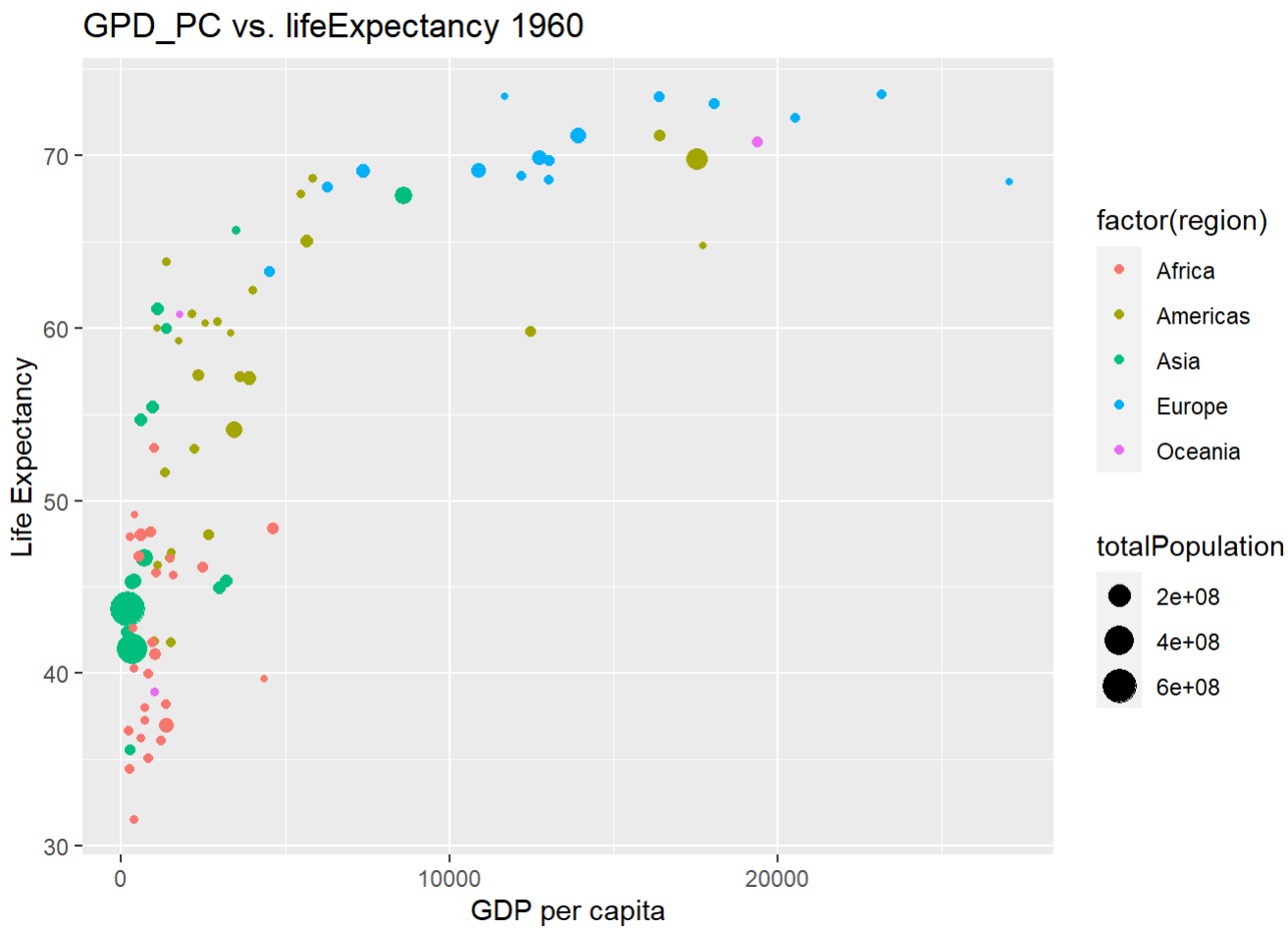
```
gapminder %>%
  filter(time == 2016) %>%
  filter(is.na(region) == FALSE) %>%
  filter(is.na(co2_PC) == FALSE) %>%
  filter(is.na(name) == FALSE) %>%
  group_by(region) %>%
  filter(rank(desc(co2_PC)) <= 3 | rank(co2_PC) <= 3) %>%
  select(name, co2_PC, region) %>%
  arrange(region)
```

```
## # A tibble: 30 × 3
## # Groups:   region [5]
##    name                                co2_PC region
##    <chr>                                <dbl> <chr>
##  1 Burundi                             0.0472 Africa
##  2 Congo, Democratic Republic of the   0.0256 Africa
##  3 Libya                               7.79   Africa
##  4 Somalia                             0.0455 Africa
##  5 Seychelles                          6.39   Africa
##  6 South Africa                        8.48   Africa
##  7 Canada                             15.1    Americas
##  8 Honduras                            1.06   Americas
##  9 Haiti                               0.275  Americas
## 10 Nicaragua                           0.887  Americas
## # … with 20 more rows
```

Problem4

Part1

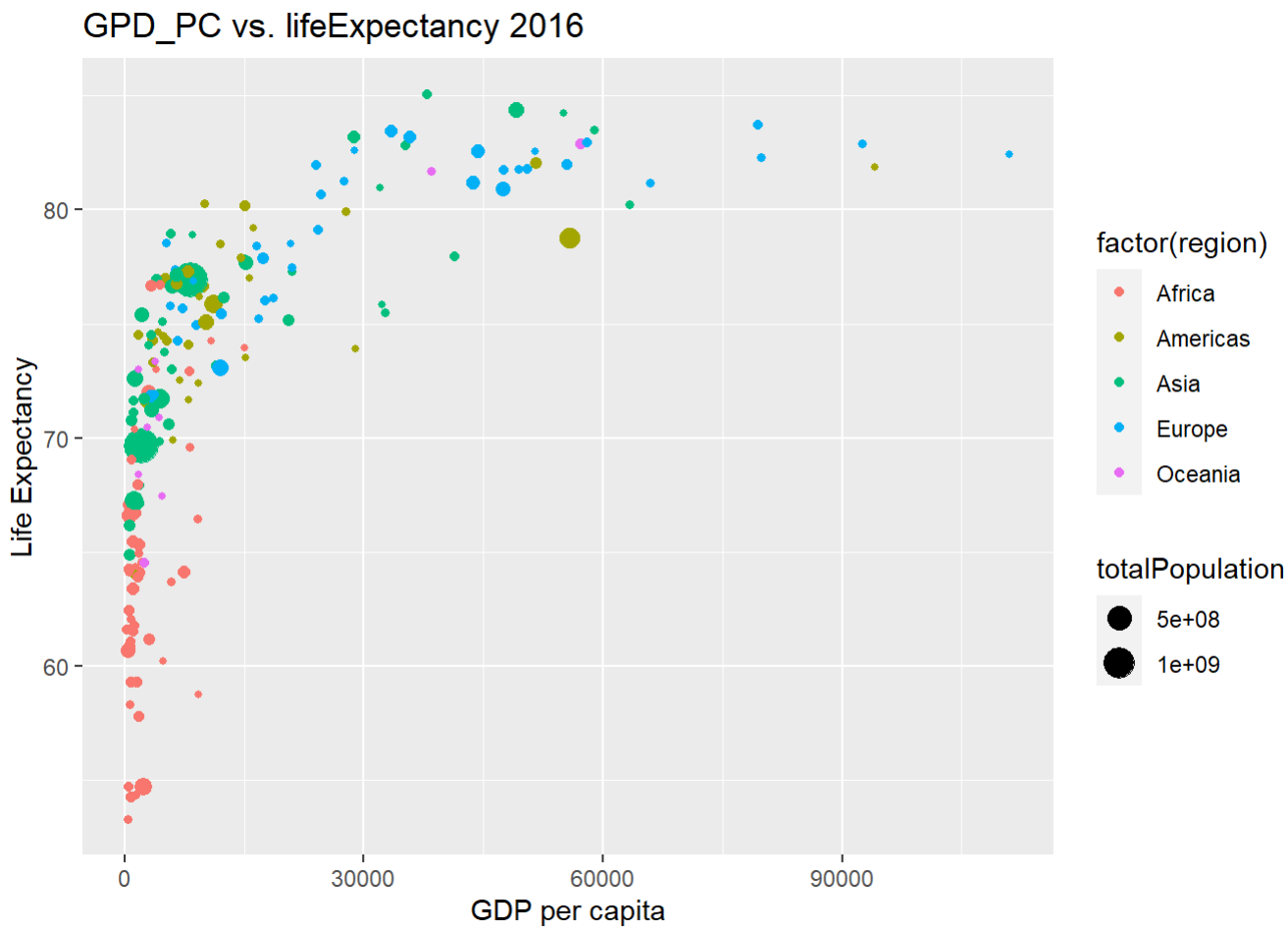
```
gapminder %>%
  filter(is.na(GDP_PC) == FALSE) %>%
  filter(is.na(lifeExpectancy) == FALSE) %>%
  filter(time == 1960) %>%
  filter(is.na(name) == FALSE) %>%
  ggplot(aes(GDP_PC,
             lifeExpectancy,
             col = factor(region),
             size = totalPopulation)) +
  geom_point() +
  labs(title = "GPD_PC vs. lifeExpectancy 1960", x = "GDP per capita", y = "Life Expectancy")
```



Based on the graph i get, the higher the GDP per capita, the higher the life expectancy will be.

Part 2

```
gapminder %>%
  filter(is.na(GDP_PC) == FALSE) %>%
  filter(is.na(lifeExpectancy) == FALSE) %>%
  filter(time == 2019) %>%
  filter(is.na(name) == FALSE) %>%
  ggplot(aes(GDP_PC,
             lifeExpectancy,
             col = factor(region),
             size = totalPopulation)) +
  geom_point() +
  labs(title = "GPD_PC vs. lifeExpectancy 2016", x = "GDP per capita", y = "Life Expectancy")
```



Part 3

Comparing these 2 plots, the overall GPD and life expectancy have increased in 2016 compare with 1960. So people can earn more money and live more happily through the last 60 years.

Part 4

```
gapminder %>%
  filter(is.na(lifeExpectancy) == FALSE) %>%
  filter(is.na(name)== FALSE)  %>%
  filter(is.na(region) == FALSE) %>%
  filter(time %in% c(1960, 2019)) %>%
  group_by(region, time) %>%
  summarize(average= mean(lifeExpectancy))
```

```
## `summarise()` has grouped output by 'region'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 10 × 3
## # Groups:   region [5]
##    region    time average
##    <chr>    <dbl>   <dbl>
##  1 Africa    1960    41.5
##  2 Africa    2019    64.1
##  3 Americas  1960    58.6
##  4 Americas  2019    75.8
##  5 Asia      1960    51.6
##  6 Asia      2019    74.6
##  7 Europe    1960    68.3
##  8 Europe    2019    79.4
##  9 Oceania   1960    56.4
## 10 Oceania   2019    73.5
```

The result fit what i see from the above 2 plots

Part 5

```
gapminder %>%
  filter(is.na(lifeExpectancy) == FALSE) %>%
  filter(is.na(name)== FALSE)  %>%
  filter(is.na(region) == FALSE) %>%
  filter(time %in% c(1960, 2019)) %>%
  group_by(region, time) %>%
  summarise(average = mean(lifeExpectancy)) %>%
  mutate(prev = lag(average, default = 0), growth = average - prev) %>%
  arrange(desc(growth))
```

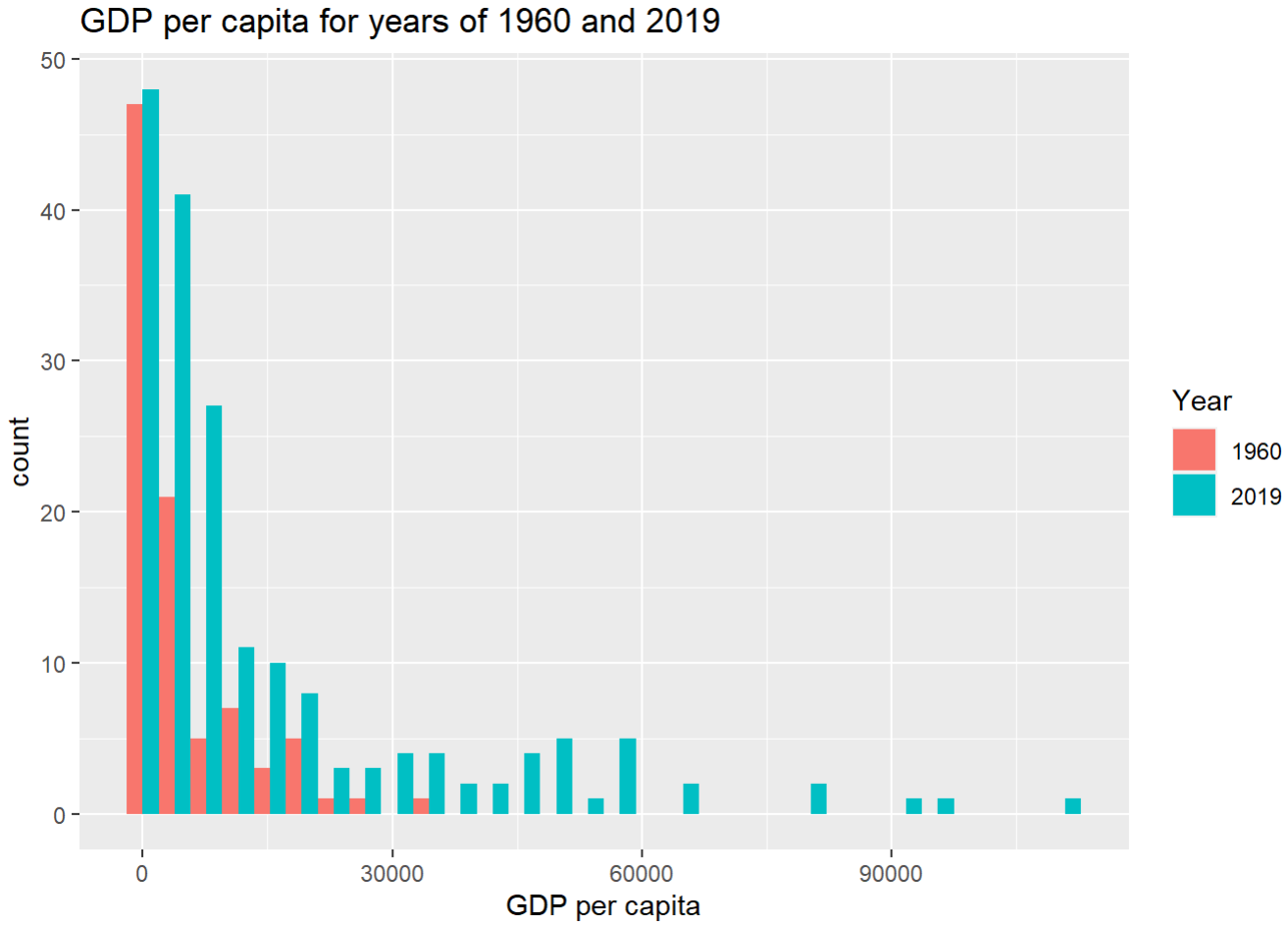
```
## `summarise()` has grouped output by 'region'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 10 × 5
## # Groups:   region [5]
##    region    time average  prev growth
##    <chr>    <dbl>   <dbl> <dbl>  <dbl>
##  1 Europe    1960    68.3   0     68.3
##  2 Americas  1960    58.6   0     58.6
##  3 Oceania   1960    56.4   0     56.4
##  4 Asia      1960    51.6   0     51.6
##  5 Africa    1960    41.5   0     41.5
##  6 Asia      2019    74.6  51.6   23.0
##  7 Africa    2019    64.1  41.5   22.6
##  8 Americas  2019    75.8  58.6   17.2
##  9 Oceania   2019    73.5  56.4   17.1
## 10 Europe    2019    79.4  68.3   11.1
```

The life expectancy for each continent have all increased over the past 60 years, and Asia and Africa are the 2 continents that increase the most.

Part 6

```
gapminder %>%
  filter(time %in% c(1960, 2019)) %>%
  filter(is.na(GDP_PC) == FALSE) %>%
  ggplot(aes(GDP_PC,
             fill = as.factor(time))) +
  geom_histogram(bins = 30, position = "dodge") +
  scale_fill_discrete("Year") +
  labs(title = "GDP per capita for years of 1960 and 2019", x = "GDP per capita", y = "count")
```



Part 7

```
gapminder %>%
  filter(time %in% c(1960, 2019)) %>%
  filter(is.na(name) == FALSE) %>%
  group_by(time) %>%
  mutate(ranking = rank(desc(lifeExpectancy))) %>%
  select(time, name, ranking) %>%
  filter(name == "United States of America")
```

```
## # A tibble: 2 × 3
## # Groups:   time [2]
##    time name                     ranking
##   <dbl> <chr>                      <dbl>
## 1  1960 United States of America      17
## 2  2019 United States of America      46
```

Part 8

```
gapminder %>%
  filter(time %in% c(1960, 2019)) %>%
  filter(is.na(name) == FALSE) %>%
  group_by(time) %>%
  mutate(ranking = rank(desc(lifeExpectancy))) %>%
  filter(is.na(lifeExpectancy) == FALSE) %>%
  mutate(relative_rank = ranking / length(unique(name))) %>%
  select(time, name, ranking, relative_rank) %>%
  filter(name == "United States of America")
```

```
## # A tibble: 2 × 4
## # Groups:   time [2]
##    time name                     ranking relative_rank
##   <dbl> <chr>                      <dbl>         <dbl>
## 1  1960 United States of America      17        0.0904
## 2  2019 United States of America      46        0.235
```

Finally

I worked on this assignment for about 5 hours